Managing Data

As we just discussed, Azure Machine Learning has two data management tools that we need to consider: **Datastores** and **datasets**. At first the distinction between the two may not be entirely clear, so let's have a closer look at what each one does and how they are related.

Datastores vs. Datasets

Datastores offer a layer of abstraction over the supported Azure storage services. They store all the information needed to connect to a particular storage service. Datastores provide an access mechanism that is independent of the computer resource that is used to drive a machine learning process.

Datasets are resources for exploring, transforming, and managing data in Azure ML. A dataset is essentially a reference that points to the data in storage. It is used to get specific data files in the datastores.

QUESTION 1 OF 2

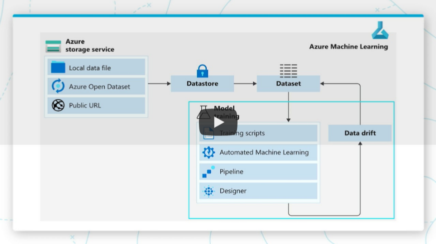
For each of the descriptions below, mark whether it refers to **datastores** or **datasets**.

Submit to check your answer choices!

DESCRIPTION	DATASTORE OR DATASET?
Answers the question, "how do I get access to specific data files?"	Dataset
Answers the question, "how do I securely connect to the data in my Azure storage?"	Datastore
Keeps connection information internal, so it is not exposed in scripts.	Datastore
Points to specific files in your underlying storage that you want to use in your ML experiments.	Dataset

SUBMIT

The Data Access Workflow



The steps of the data access workflow are:

1. **Create a datastore** so that you can access storage services in Azure.
2. **Create a dataset**, which you will subsequently use for model training in your machine learning experiment.
3. **Create a dataset monitor** to detect issues in the data, such as data drift.

In the video, we mentioned the concept of *data drift*. Over time, the input data that you are feeding into your model is likely to change—and this is what we mean by **data drift**. Data drift can be problematic for model accuracy. Since you trained the model on a certain set of data, it can become increasingly inaccurate and the data changes more and more over time. For example, if you train a model to detect spam in email, it may become less accurate as new types of spam arise that are different from the spam on which the model was trained.

As we noted in the video, you can set up *dataset monitors* to detect data drift and other issues in your data. When data drift is detected, you can have the system automatically update the input dataset so that you can retrain the model and maintain its accuracy.

QUESTION 2 OF 2

Below are the main processes of the data access workflow. Can you match each process with the correct step?

Mount dataset to experiment compute target

STEP	PROCESS
Step 1	Create a datastore
Step 2	Create a dataset
Step 3	Create a data monitor

SUBMIT

NEXT